Assessing the Reliability and Validity of Expert Interviews

Dorussen, Han; Lenz, Hartmut; Blavoukos, Spyros

Postprint / Postprint
Zeitschriftenartikel / journal article

European Union Politics
DOI: 10.1177/1465116505054835
Volume 6 (3): 315–337

SAGE Publications
London, Thousand Oaks CA,
New Delhi

Assessing the Reliability and Validity of Expert Interviews

Han Dorussen
University of Essex, UK

Hartmut Lenz
University of Essex, UK

Spyros Blavoukos
University of Essex, UK

ABSTRACT

Testing the reliability of experts should be a key element of expert interviews. Using the Condorcet Jury Theorem, it is shown that expert reliability can provide an indication of the validity of expert-opinion data. The theoretical framework is applied to expert-interview data collected in the Domestic Structures and European Integration (DOSEI) project. Special attention is paid to the role of 'leading' experts and salient issues. Evaluating the DOSEI data, the main findings are that (i) with some exceptions, there are acceptable levels of inter-expert agreement, (ii) whether the leading expert is included or not does not make a large difference to expert agreement, and (iii) experts are more in agreement on salient issues.

KEY WORDS

- Condorcet Jury Theorem
- European Constitution
- expert interviews
- inter-coder agreement
- reliability
- saliency of issues

Introduction

When asking for directions, is it better to ask one or several persons? If they happen to agree, one may feel more confident. Then again, what to do if they give differing (possibly even conflicting) advice? If one comes across a police officer or taxi driver, then it is probably best simply to follow their instructions. Authoritative guidance carries even more weight when asking the way for a less obvious landmark – for example, a particular street rather than a central station.

Likewise, we need to evaluate the validity and reliability of the information provided by the interviewees in survey research and expert interviews. In mass surveys we collect a large number of data points, but we may doubt the quality of the individual responses because the interviewees are generally poorly informed and motivated. In contrast, experts should be better informed and more motivated, but we generally rely on only a few data points. At the same time, we may suspect that not all of our experts are equally knowledgeable and, of course, even experts occasionally make mistakes.

In this article, we address these long-standing issues as they apply specifically to expert-opinion data collected in the Domestic Structures and European Integration (DOSEI) project. The DOSEI project aims to determine the position of the EU members (plus the Commission and European Parliament) on the draft Constitution for the European Union. As part of the DOSEI project, country experts were asked to assess the national position on the draft Constitution. A total of 77 experts were interviewed, varying from one to six experts for any particular political actor.

Our first research question is how coherent the experts are in identifying the policy positions. The DOSEI expert interviews contain two – rather unusual – features that are especially useful for our research purposes. First of all, the interviewers were asked to provide a personal assessment of the quality of the various experts. As with asking for directions, face-value validity often depends on intangibles such as body language or the provision of additional, not directly relevant, details. On the basis of this additional information, DOSEI identified a 'leading expert' for each country.

A further research question is whether we can corroborate the quality of the 'leading expert' on the basis of a more objective evaluation of the answers provided by the various experts. Is the coherence between the leading and other experts greater than the coherence of the non-leading experts?

The second feature is that, as part of the DOSEI interviews, experts were asked to identify the most salient issues for 'their' country. Our final research question is whether the *salience* of an issue affects expert coherence. To use the

metaphor of asking directions one last time, just as it is reasonable to expect more consistent instructions when trying to get to a prominent or generally familiar spot, we predict higher expert coherence for more salient issues.

Expert interviews are an attractive data collection method, because they allow researchers to bridge the divide between case studies and the comparison of a large number of countries based on more general and publicly available data. Further, expert interviews give the researchers control over the dimensions that are central to the comparative research.[1] Consequently, a clear theoretical framework can be used to facilitate rigorous comparisons.

The use of expert interviews as an instrument to collect data is quite common in political science and, in particular, in studies of the European Union.[2] Experts may provide a unique source for 'inside' information about the policy-making process. In political science, experts 'code' information about policy processes and political actors. They are thus comparable to doctors diagnosing patients or coders of texts in content analysis. In medicine, management and marketing, communication and education studies, and linguistics, the use of multiple experts/coders and subsequent reporting of inter-expert/coder reliability is commonplace. However, a brief survey on the use of expert-opinion data in European Union policy studies found that the issue of inter-expert reliability is largely ignored.

The primary research interest in this article is to evaluate the quality of the expert-opinion data collected as part of the DOSEI project. In order to do so, two more general questions need to be considered. First, does the reliability of expert-opinion data provide an indication of their validity? The Condorcet Jury Theorem is used as the theoretical framework to discuss this issue. Second, how can we best determine the reliability of expert-opinion data? In fact, there exists a multitude of indices for inter-rater or inter-coder agreement. Most of the statistics available have been developed for the needs of specific disciplines. Consequently, they reflect the requirements of a particular discipline and their applicability to other research areas may be limited. We will discuss the strengths and weaknesses of the most commonly used indices of inter-coder agreement: percent agreement, Cohen's κ, and Krippendorff's α. The evaluation of the DOSEI data relies on the same indices.

In the remainder of this article we first discuss various theoretical arguments for the use of multiple experts to increase the validity of the data, and consider the relation between validity and reliability. We also provide a brief survey of the debate surrounding the various indicators for inter-coder reliability. We conclude that there are good reasons for consulting multiple experts, whenever possible. Moreover, it is important to report inter-expert reliability. Even though existing indices of inter-coder reliability have their shortcomings, they provide valuable information and are easily accessible.

The subsequent sections evaluate the inter-coder reliability of the DOSEI data. Special attention is paid to the role of 'leading' experts and salient issues. We find that (i) with some exceptions, the inter-expert agreement in the DOSEI data is at an acceptable level, (ii) whether the leading expert is included or not does not make a large difference to expert agreement, and (iii) experts are more in agreement on salient issues.

Inter-coder reliability and validity

It is important not to confuse the reliability of data and the validity of research results based on them. Reliability *sets limits to the potential* validity of research results, but reliability does not *guarantee* the validity of research results. Reliability is a necessary but not a sufficient condition for validity (King et al., 1994: 24; Krippendorff, 2004b: 212–13). Nevertheless, it makes intuitive sense to use agreement among *experts* as an indication of the quality of the information they provided. The Condorcet Jury Theorem (CJT), moreover, provides a possible theoretical justification.[3] In its most basic form, the CJT applies to a group of individuals who independently have to make a binary decision that is either 'right' or 'wrong', and where each individual has a fixed probability of being right. Condorcet wanted to know the probability of the majority being right under these assumptions.

The binomial probability formula can be used to find the answer to Condorcet's problem. Assume that the individual probability of being right equals p. In a group of n individuals, the probability of a majority providing the right answer (P) then is the sum of $(n + 1)/2$ to all n individuals being right, or

$$P = \sum_{x=(n+1)/2}^{n} \binom{n}{x} p^x (1-p)^{n-x} \tag{1}$$

If all individuals are equally competent and more likely to be right than wrong (or $p > .5$), equation (1) can be applied to demonstrate the two parts of the CJT. First, since

$$\lim_{n\to\infty} \sum_{x=(n+1)/2}^{n} \binom{n}{x} p^x (1-p)^{n-x} = 1, \tag{2}$$

it becomes extremely likely that the majority is right when the number of individuals increases. This is known as the asymptotic part of the theorem. The non-asymptotic part of the Jury Theorem holds that the majority is more reliable than each individual citizen. The second part of the theorem can be proven by showing that

$$p < \sum_{x=(n+1)/2}^{n} \binom{n}{x} p^x (1-p)^{n-x} \tag{3}$$

for groups of any size n (Ben-Yashar and Paroush, 2000).

For our purposes, there are two important extensions to the original CJT. The first deals with the justification of supermajorities, and the second relaxes the assumption of identical (or homogeneous) competence. A supermajority voting rule requires a fraction $q > (n + 1)/2$ to choose an alternative. Ben-Yashar and Paroush (2000) demonstrate that requiring a larger majority increases the probability of a 'right' choice. In other words, not only does agreement among the majority indicate that they are more likely right than wrong, a larger majority is even more likely to be right (at least if we hold the number of individuals constant).

The CJT has been further generalized by relaxing the assumption that all individuals have the same level of competence. Instead we can allow for heterogeneous levels of competence, where p_i stands for the probability that any particular individual i is right. Consequently, $\bar{p} = \frac{1}{n}\sum_i p_i$ becomes the average probability of a group of individuals being right. It is straightforward to generalize the asymptotic part of the Jury Theory using average competence; if $.5 < \bar{p} < .1$, and $n > 2$, then $P > \bar{p}$ and P approaches 1 as n goes to infinity (Owen et al., 1989).

Given varying levels of competence, it is easy to construct examples in which the most competent individual is more competent than the majority. Borrowing the example provided in Nurmi (2002: 52), consider a group of three individuals with $p_1 = .9$, $p_2 = .7$, and $p_3 = .6$. In this case, $P = .834$ and $\bar{p} = .73$. Although the majority is still more likely to be right than the average individual ($P > \bar{p}$), the most competent individual is more competent than the majority ($p_1 > P$). Ben-Yashar and Paroush (2000) demonstrate that the first part of the above statement holds in general. Nitzan and Paroush (1982) specify the condition under which the second part of the statement can occur: given an odd number of n individuals with varying competence but for all, $p_i > .5$, and labelled in non-increasing order of competence, i.e. $p_i \geq p_j$ if $i < j$, then the non-asymptotic part of the CJT does *not* hold, if

$$\frac{p_1}{1-p_1} > \prod_{i=2}^{n} \frac{p_i}{1-p_i}. \tag{4}$$

To put it somewhat differently, suppose a jury of three individuals. The most competent member has $p_1 = .9$, and the other two members are equally but less competent, or $p_1 > p_2 = p3$. In this case, the majority is less competent than the most competent member if $p_2 = p_3 < .75$.

The CJT is directly relevant for our main research questions. If we are willing to accept the assumption that each expert is more likely to provide the right rather than a wrong assessment of an actor's policy position, then the majority is more likely to be correct. With respect to our first research question on the importance of expert coherence, the CJT demonstrates that, the more the experts agree, the more likely it is that they are right, assuming that $\bar{p} > .5$. Moreover, we can generally have more confidence in the majority opinion if we increase the number of experts. Even if we suspect that the experts vary in their competence, the probability of the majority of experts being right often (but not always) exceeds the probability of the most competent expert being right. With respect to our second research question, if the 'leading' expert is indeed more likely to be correct, she should increase the average competence of the experts. Consequently, inclusion of the 'leading' expert should improve the validity of the majority agreement. Concerning our third research question, if experts are on average more competent on salient issues, the CJT demonstrates that the majority opinion is more likely to be valid.

The conclusions with respect to the second and third questions can be stated as hypotheses guiding our empirical assessment of expert coherence in the DOSEI project.

H1: Agreement within a group of experts including the 'leading' expert should be higher than that within a group excluding the 'leading' expert.

H2: Expert agreement on salient issues should be higher than expert agreement on non-salient issues.

Condorcet Jury Theorem and observed agreement

In the CJT the expected probability of agreement (P) is measured by way of equation (2), i.e. as the sum of the probabilities that a majority of individuals – from $(n + 1)/2$ to n, where $n > 2$ – agree on the 'right' answer. In contrast, inter-coder indices of agreement are based on observed agreement as a proportion of the times coders could agree. The latter indicator of agreement – commonly referred to as ($P(A)$) – differs from Condorcet agreement in three respects: (a) $P(A)$ evaluates agreement between pairs of experts as a proportion of the maximum possible number of pairs that could have agreed, (b) $P(A)$ includes agreement on all possible categories, including 'right' as well as 'wrong' answers, and (c) $P(A)$ evaluates agreement for two or more experts on two or more answer categories. Following Siegel and Castellan (1988: 286), $P(A)$ can be defined as:

$$P(A) = \left[\frac{1}{Sn(n-1)} \sum_{i=1}^{S} \sum_{j=1}^{m} c_{ij}^2 \right] - \frac{1}{n-1}, \quad (5)$$

where S is the number of subjects, m the number of categories, n the number of experts, and cij the number of experts who assign a subject i as an instance of j.

However, if the assumptions of the CJT apply, observed agreement can be used as an indicator of majority agreement. In terms of equation (5), the CJT assumes that $m = 2$ and that the probability of each expert coding 'right' is $p_i > .5$. Consequently, the probability that two experts, i and j, independently both code 'right' equals $p_i \times p_j$ and the probability that they both code 'wrong' equals $(1 - p_i) \times (1 - p_j)$. Finally, the probability of disagreement equals $p_i \times (1 - p_j) + p_j \times (1 - p_i)$. However, because $p_i, p_j > .5$ (by assumption), it follows that $p_i \times p_j > (1 - p_i) \times (1 - p_j)$ and $p_i \times p_j + (1 - p_i) \times (1 - p_j) > p_i \times (1 - p_j) + p_j \times (1 - p_i)$. In words, if the experts are more competent, we expect more agreement on the 'right' rather than the 'wrong' classification, and we expect to observe a higher proportion of pairs of experts in agreement.

It should be obvious that the assumptions about the competence of the experts are crucial for these conclusions. If one of a pair of experts codifies at random, $p_i = .5$, agreement and disagreement are equally likely. Even worse, if experts are biased towards the 'wrong' answer, $p_i < .5$, we still observe more agreement than disagreement, but this time the experts will agree on the 'wrong' classification. If the number of categories or alternatives increases, the assumption that an expert is more likely to identify the 'right' rather than one of the 'wrong' alternatives becomes less reasonable. For example, when asking an expert to identify a position on a dimension (a subset of real numbers), the probability that exactly the 'right' position is identified approaches zero, and the probability of two experts agreeing completely is also very close to zero. Clearly, in order for the CJT to apply, the number of possible answer categories has to be fairly small. In the DOSEI project, the experts were asked to identify actors' ideal policy positions, but the actual questions made the experts choose between two and five possible answer categories (see Table 2 below). The assumptions of the CJT are thus reasonable for the DOSEI data.

Measuring agreement and inferring reliability

The assessment of inter-coder reliability is a two-stage process. First, two or more coders independently categorize the research units of interest (patients, texts, etc.). Second, a numerical index is calculated based on the

categorizations, quantifying the extent of agreement among the multiple coders. Reliability is inferred from the numerical index according to its particular assumptions. A large number of indices are available. Broadly, they can be distinguished as either extensions of percent agreement analysis (Holsti, 1969) or correcting the observed agreement for chance agreement (Scott, 1955; Fleiss, 1971; Krippendorff, 2004b).

Percent agreement is a simple index that is most commonly calculated as the percentage of all coding decisions on which the coders agree. In equation (5) above, percent agreement is the index of observed agreement $P(A)$. The basic advantages of this index are that it is easy to calculate and it allows for any number of coders. However, without any correction, percent agreement becomes (probabilistically) increasingly more likely the fewer categories are available for coding.

Chance-corrected agreement indices distinguish between $P(A)$, the observed agreement between coders, and $P(E)$, the probability of agreement among coders owing to chance. In general, they can all be expressed as:

$$\frac{P(A)-P(E)}{1-P(E)}. \tag{6}$$

The correction for chance is important for enabling observers to judge the reliability of the data. However, the various agreement indices differ in their correction for chance agreement. Cohen's κ is based on each coder's personal distribution (Cohen, 1960; Di Eugenio and Glass, 2004). The Krippendorff α (as well as other indices by Scott, 1955; Fleiss, 1971; Siegel and Castellan, 1988) assumes one distribution for all coders, which is derived from the total proportion of categories assigned by all coders. Table 1 provides an overview of the various chance-corrected indices for agreement.

Even though the differences between these indices have been theoretically acknowledged, in most empirical research they have been neglected – in practice the indices produce often similar results. Moreover, as Table 1 shows, the indices converge if the number of subjects (S) grows large. Di Eugenio and Glass (2004) notice, however, that the different computations of $P(E)$ have a larger effect given smaller values of κ.

There are some further limitations to the κ and α that are particularly germane to our research purposes. Both statistics are affected by skewed distributions of categories, or the *prevalence* problem, and perform best if the data are less skewed (Di Eugenio and Glass, 2004; Krippendorff, 2004a, b). This is especially important for nominal data, where the categories are chosen more or less arbitrarily. Even more problematic is that Cohen's κ is affected by the degree to which the coders disagree (the *bias* problem). Actually, it is relatively easy to show that some indices increase if the coders are *less* similar.[4]

Table 1 Inter-coder agreement measures

Percent agreement ($P(A)$)	$P(A)=\left[\frac{1}{Sn(n-1)}\sum_{i=1}^{S}\sum_{j=1}^{m}c_{ij}^2\right]-\frac{1}{n-1}$
Cohen's κ	$\frac{P(A)-P(E)}{1-P(E)}$, where $P(E)=\sum_{j=1}^{m}(p_{j1}\times p_{j2})$
Fleiss–Siegel & Castellan κ	$\frac{P(A)-P(E)}{1-P(E)}$, where $P(E)=\sum_{j=1}^{m}\left(\frac{\sum_{i=1}^{S}c_{ij}}{nS}\right)^2$
Krippendorff's α	$\frac{P(A)-P(E)}{1-P(E)}$, where $P(E)=\sum_{j=1}^{m}\left(\frac{\sum_{i=1}^{S}c_{ij}}{nS}\right)\left(\frac{\left(\sum_{i=1}^{S}c_{ij}\right)-1}{nS-1}\right)$

Note: S is the number of classifications, n is the number of raters, m is the number of categories, c_{ij} is the number of assignments, p_{j1} is the marginal probability of choosing category j for rater 1. Note that Cohen's κ is defined only for two raters.

None of the chance corrections is entirely intuitive from the perspective of the Condorcet Jury Theorem. Cohen's κ quantifies the deviation of observed agreement from statistical independence. The Fleiss–Siegel and Castellan κ and Krippendorff α correct observed agreement for the likelihood that the matched observations are chance events. Based on the CJT, however, the appropriate benchmark for expected agreement is randomness of classification rather than statistical independence or random matching.

There is no consensus on the appropriateness of the various indices across various disciplines. In the marketing, advertising and consumer behaviour literature, simple percent agreement is the most commonly used index for inter-coder reliability (Hughes and Garrett, 1990); Krippendorff's and Holsti's methods of assessment follow at a distance (Kolbe and Burnett, 1991). In the computational linguistics literature, Cohen's κ coefficient is considered the de facto standard for inter-coder agreement despite some criticism (Carletta, 1996; Di Eugenio and Glass, 2004). The same conclusion was reached in content analyses of news- and media-related articles (Riffe and Freitag, 1997; Lombard et al., 2004).

Scholars disagree about what constitutes an acceptable level of reliability in the various indices. In general, indices based solely on percent agreement provide systematically higher values. Hence, higher critical values should be adopted for these indices. Chance-corrected indices are considered more

conservative in their estimates and should be controlled for lower critical values. The various chance-corrected indices (the κ and α statistics) vary in how they are affected by skewed distribution of categories (the *prevalence* problem) rather than disagreement of coders (*bias* problem). This makes it nearly impossible to select critical values for reliability that apply across indices or even across research designs (Krippendorff, 2004a: 429).

As a rule of thumb, Neuendorf (2002) proposes that, for percent agreement, a coefficient of .9 or greater always indicates high reliability and a coefficient of .8 or greater is acceptable in most situations; below that level, acceptance or rejection depends on the nature of the study and the intentions of the researcher. Krippendorff (2004b) considers it customary to require $\alpha \geq .8$, or, if tentative conclusions can be accepted, $\alpha \geq .667$ as the lowest conceivable limit. Ultimately, the choice of an acceptable threshold for agreement involves a subjective evaluation balancing the risks of invalid data against losing the data completely.

Despite all these concerns, there are important areas of agreement. First of all, in research that relies on subjective evaluation, it is important to provide a sense of the reliability of the coders. In this respect, it is worrying that reliability tests are rare in political science, even though content analysis and expert interviews are commonly used. Second, the minimum requirements to produce meaningful reliability results are the use of a representative set of units for the testing and independent codification by all coders of the set of units under the same conditions. Third, there is a consensus that agreement indices based on association, correlation and consistency coefficients are inappropriate for inferring reliability from agreement.

Coherence in the DOSEI expert interviews

We apply the various indices of inter-coder agreement to the data collected as part of the Domestic Structures and European Integration (DOSEI) project. The DOSEI project has used an expert survey to collect data on the positions of the EU members (plus the Commission and European Parliament) regarding the draft Constitution for the European Union. Country experts were asked about the national position – the position of the government – on 24 issues central to the draft Constitution. The survey used a total of 50 questions. Table 2 shows the distribution of the number of answer categories for the different questions (both including and excluding sub-issues). Since the number of answer categories is always reasonably small (never more than five possible categories), the DOSEI data satisfy the assumptions underlying the Condorcet Jury Theorem as well as the various inter-coder agreement indices.

Table 2 Number of answer categories per question

Number of answer categories	*Number of questions, excluding sub-questions*	*Number of questions, including sub-questions*
2	4	21
3	6	8
4	6	18
5	3	3
Average	3.61	3.06
Total number of questions	19	50

The DOSEI project has selected the experts carefully. In the first instance, academics were asked to identify central actors in the policy formation process. Next, these central actors were asked to identify people they considered most knowledgeable. The people who were mentioned most frequently were eventually approached to serve as experts. Great care was taken to find people who were as close as possible to the policy formation process for the Intergovernmental Conference on the European Constitution. Whenever several experts were approached, it was deemed important that they occupy positions in different institutions relevant in the policy formation process. The expert interviews all took place in the autumn of 2003 as close as possible to (but before) the Intergovernmental Conference on the European Constitution held in December 2003.

Table 3 reveals the agreement of the experts in the DOSEI project. For two countries (Cyprus and Greece) only one expert was interviewed, which makes it impossible to assess any inter-expert coherence; we therefore ignore these cases for the rest of the article. For the remaining 25 political actors, a total of 75 experts were consulted. In 12 cases, only two experts were interviewed. Information for the other 13 actors was collected using up to six experts. Table 3 lists three indices of inter-expert agreement: percent agreement, κ and α. We find that κ and α are highly correlated, but there are some differences with respect to percent agreement.

The various indicators suggest moderate agreement among the experts in identifying the national positions. Based on the κ and α statistics, agreement is particularly poor in the case of Hungary. Percent agreement is, however, notably higher for Hungary, although is still the lowest compared with all other countries. Percent agreement is also notably higher than the κ and α statistics in the cases of Portugal, Sweden, the Netherlands, Latvia and Slovenia. Inter-expert agreement is particularly strong in Luxembourg, Germany and the EU Commission and Parliament. A comparison of expert

Table 3 Indicators of expert agreement in the DOSEI project

Number of experts	*Actor/country*	*Abbrev.*	*P(A)*	κ	α	*Status of actor/country*
1	Cyprus	CYP	–	–	–	Acceding
	Greece	GRE	–	–	–	EU15
2	Commission	COM	.92	.77	.76	EU institution
	Czech Republic	CZR	.64	.41	.42	Acceding
	Estonia	EST	.79	.64	.65	Acceding
	France	FRA	.84	.69	.70	EU15
	Hungary	HUN	.59	.20	.23	Acceding
	Italy	ITA	.67	.52	.51	EU15
	Luxembourg	LUX	.93	.84	.84	EU15
	Malta	MAL	.65	.47	.44	Acceding
	Poland	POL	.81	.61	.61	Acceding
	Slovakia	SLK	.84	.72	.70	Acceding
	Spain	SPA	.83	.69	.69	EU15
	United Kingdom	UNK	.70	.49	.49	EU15
	Average coherence 2 experts (n = 12)		.78	.59	.59	
3	Belgium	BEL	.92	.77	.77	EU15
	Denmark	DEN	.85	.73	.75	EU15
	Portugal	POR	.70	.48	.47	EU15
	Sweden	SWE	.86	.53	.54	EU15
	Average coherence 3 experts (n = 4)		.83	.63	.63	
4	Austria	AUS	.80	.57	.59	EU15
	Finland	FIN	.82	.58	.58	EU15
	Germany	GER	.96	.89	.88	EU15
	Ireland	IRE	.86	.74	.74	EU15
	Lithuania	LIT	.79	.56	.57	Acceding
	Netherlands	NET	.76	.40	.40	EU15
	Average coherence 4 experts (n = 6)		.83	.62	.63	
5	Latvia	LAT	.75	.54	.55	Acceding
	Slovenia	SLN	.83	.56	.54	Acceding
	Average coherence 5 experts (n = 2)		.79	.66	.54	
6	EU Parliament	PAR	.89	.66	.66	EU institution
Average coherence EU15 (n = 14)			.82	.64	.64	
Average coherence acceding (n = 9)			.74	.52	.52	
Average coherence COM/PAR (n = 2)			.91	.71	.71	

agreement for small/medium countries with large countries (France, Germany, Italy and the United Kingdom) shows that agreement is somewhat higher for the latter than the former (average α's are .64 versus .58). The result remains constant if Spain and Poland are considered large countries as well.[5]

On average, inter-expert coherence in the original members of the European Union (EU15) is higher than in the acceding countries. The uneven distribution of EU15 and acceding countries with respect to the number of experts consulted makes it difficult to assess the effect of the number of experts on their coherence. Pearson correlations between the number of experts and the various indices for inter-expert coherence were all highly insignificant. Correlation between the number of experts and percent agreement was .29 (significance level: .16), with κ: .06 (significance level: .78) and α: .06 (significance level: .77).

Clearly, the identity of the actor matters more for expert coherence than the number of experts. As one would expect, experts are more in agreement if an actor has well-defined – and well-publicized – opinions on EU issues. The strong coherence of the experts in their evaluation of the policy positions of the European institutions is thus rather unsurprising. The rather poor coherence of the experts in some of the acceding countries resembles the opinion of the citizens in these countries that they are poorly informed about the EU draft Constitution (Flash Eurobarometer, 2004: 3–5). The same even applies to some of the original EU members. For example, expert coherence is remarkably low in the Netherlands, but mirrors the fact that less than a quarter of Dutch citizens feel well informed about the EU Constitution. We do not want to imply a causal link between public awareness and expert knowledge. It is more likely that the lack of a clear stance on several issues in some countries resulted in low agreement among experts and a perceived lack of information among the general public.

'Leading' experts and expert coherence

Obviously, researchers may question the validity of (some of) the answers provided even when experts are chosen with great care. Confidence in an expert may be based on the position or reputation of the expert, or simply on his or her behaviour during the interview. In the DOSEI project, the interviewers were asked to provide a personal assessment of the 'quality' of the various experts, and a 'leading' expert was identified for each country/actor.[6] We are, first of all, interested in whether 'leading' experts do indeed improve the consistency of the experts. In other words, are the other experts more likely to agree with the 'leading' expert than among themselves? It is, however, also interesting to see whether the 'objective' information on expert

coherence suggests the same expert as 'leading' as does interviewers' intuition.

In Figures 1, 2 and 3, we compare the inter-expert coherence (using P(A), κ and α) including and excluding the 'leading' experts. Obviously, in order to calculate expert coherence without the 'leading' expert, we need to have at least three experts initially. Consequently, for this part of the analysis, we had to ignore the 12 cases with only two experts – 14 including Greece and Cyprus, where only one expert was interviewed.

The first conclusion has to be that the 'leading' expert does not necessarily increase inter-expert coherence. In somewhat more than half of the 13 remaining cases, coherence including the 'leading' expert is about equal to or

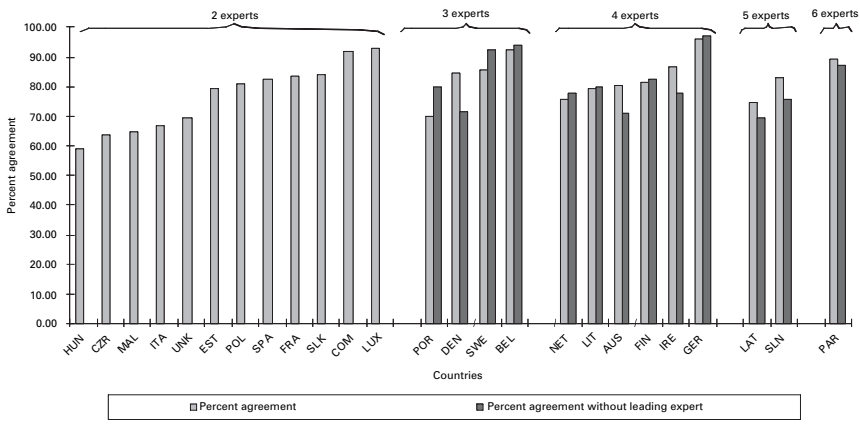

Figure 1 Comparison of expert coherence, with and without leading expert: Percent agreement.

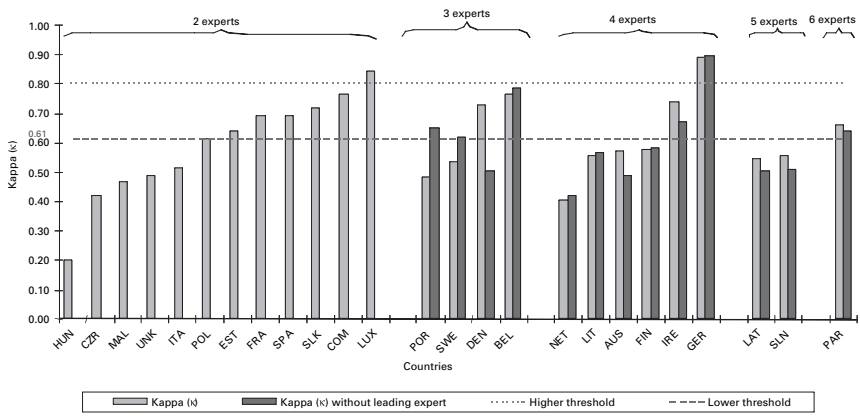

Figure 2 Comparison of expert coherence, with and without leading expert: κ.

slightly above inter-expert coherence with the 'leading' expert excluded. In 5 cases, the non-leading experts notably deviate from the 'leading' expert. Excluding the 'leading' expert clearly lowers expert coherence in three cases: Denmark, Austria and Ireland. In the Swedish and Portuguese cases, coherence among the non-leading experts is quite a bit higher than the inter-expert coherence including the 'leading' expert. By coincidence, we personally conducted the interviews in Sweden and Portugal. We remain confident that the 'leading' expert was best positioned to have access to high-quality information. One possibility is that the non-leading experts agreed on incorrect positions. It would be more worrying if the 'leading' expert misrepresented policy positions for strategic reasons (on how strategic behaviour undermines the Condorcet Jury Theorem, see Austin-Smith and Banks, 1996).

The results presented in Table 4 largely confirm this pattern. Using the α statistic, we have calculated the average of the pair-wise agreement of each expert with the others. This allows us to compare the average pair-wise agreement of the 'leading' expert (with the other experts) with that of each of the non-leading experts (with the other experts, including the 'leading' expert). Of course, this is only possible if three or more experts were originally consulted. The somewhat remarkable finding is that in only about half of all cases was the 'leading' expert most in agreement with the other experts. In seven cases, there was more agreement with one of the non-leading experts – most strongly in Portugal, the Netherlands and Lithuania.

The overall finding is that the 'leading' expert has only a minor impact on inter-expert coherence. Thus, there is little support for Hypothesis 1 that there is most agreement with the 'leading' expert. In only three cases does

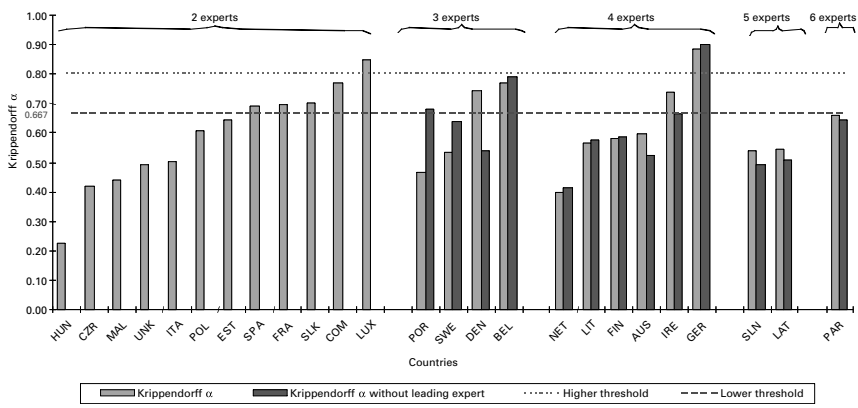

Figure 3 Comparison of expert coherence, with and without leading expert: Krippendorff α.

Table 4 Comparison of average pair-wise agreement between each expert and all other experts (Krippendorff α)

Country	*Abbrev.*	*Leading expert*	*2nd expert*	*3rd expert*	*4th expert*	*5th expert*	*6th expert*
Portugal	POR	.49	.68	.63			
Sweden	SWE	.51	.59	.56			
Belgium	BEL	.75	.78	.76			
Denmark	DEN	.76	.75	.55			
Netherlands	NET	.39	.47	.44	.35		
Lithuania	LIT	.54	.65	.54	.51		
Finland	FIN	.58	.61	.61	.53		
Austria	AUS	.61	.60	.59	.43		
Ireland	IRE	.82	.79	.77	.62		
Germany	GER	.90	.91	.91	.86		
Latvia	LAT	.59	.56	.55	.52	.41	
Slovenia	SLN	.66	.59	.57	.55	.53	
Parliament	PAR	.73	.71	.69	.68	.66	.53

Notes: 2nd to 6th experts are listed in decreasing order. Highest average pair-wise coherences are underlined. Only countries with three or more experts are listed.

the exclusion of the 'leading' expert conform to the prediction of the Hypothesis 1 of decreased coherence. Moreover, the analysis of pair-wise coherence shows most agreement with the 'leading' expert in only about half of all cases.

Issue-salience and expert coherence

As the final question of the interview, the experts were asked to identify the key issues for their actor in the Intergovernmental Conference on the European Constitution. This question allows us to determine what the most salient issues are, at least according to the experts. Table 5 simply lists the proportion of issues that were considered 'salient'. The proportion of salient issues ranges from about one-tenth (Belgium) to approximately one-third (France and Germany). Comparing the original EU15 with the acceding states or the European institutions, we find only small differences in the average proportion of salient issues.

The expectation is to find higher inter-expert coherence for the salient issues (Hypothesis 2). In Figures 4 and 5, we compare the percent agreement and Krippendorff α values for the salient issues with the original α's (Table 3 and Figures 1 and 3).[7] In strong support of Hypothesis 2, inter-expert coherence is indeed higher for the salient issue in 20 of the 25 cases; in 14 cases, it

Table 5 Prevalence of salient issues in EU member states

Country	*Abbrev.*	*Key issues as % of total number of issues*
Belgium	BEL	10.00
Italy	ITA	17.14
Netherlands	NET	17.14
Spain	SPA	17.14
Portugal	POR	18.57
Austria	AUS	20.00
United Kingdom	UNK	20.00
Denmark	DEN	22.86
Finland	FIN	24.29
Luxembourg	LUX	27.14
Sweden	SWE	27.14
Ireland	IRE	31.43
France	FRA	32.86
Germany	GER	34.29
Average EU15 (n = 14)		22.86
Poland	POL	12.86
Malta	MAL	14.29
Lithuania	LIT	15.71
Czech Republic	CZR	18.57
Latvia	LAT	18.57
Slovenia	SLN	20.00
Hungary	HUN	21.43
Slovakia	SLK	27.14
Estonia	EST	31.43
Average acceding (n = 9)		20.00
Parliament	PAR	20.00
Commission	COM	25.71
Average EU institution (n = 2)		22.86

Notes: Greece and Cyprus excluded. An issue is considered salient if at least one expert mentions the issue as a key issue.

is substantially higher. In 5 cases, the coherence is slightly lower for the salient issues, but generally only marginally so. We find notably lower expert coherence for the salient issues only in Poland and Belgium. These two countries also have the lowest proportion of salient issues. In the case of Poland, the government position on some of the key issues, in particular the distribution of vote shares, had been widely discussed and publicized. It is thus

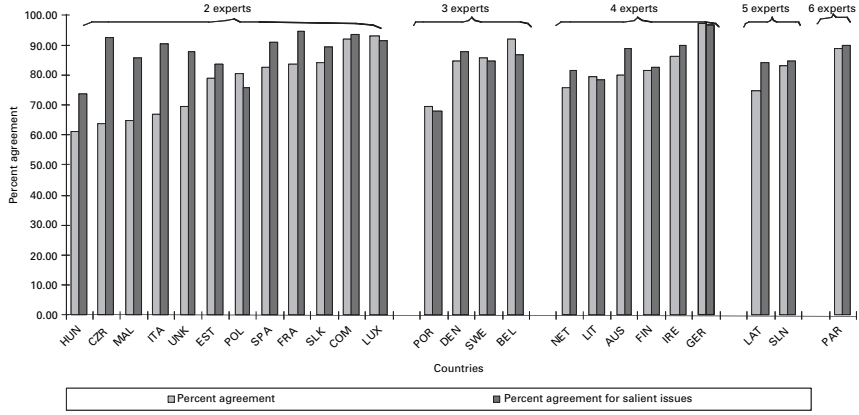

Figure 4 Comparison of expert coherence, all issues versus salient issues: Percent agreement.

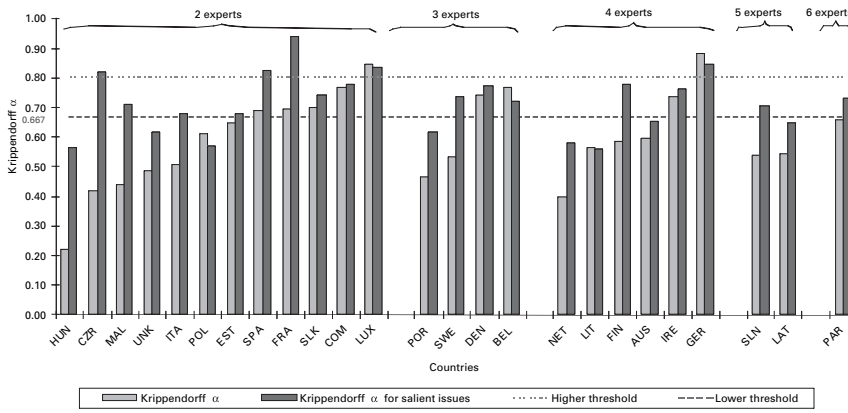

Figure 5 Comparison of expert coherence, all issues versus non-salient issues: Krippendorff α.

especially surprising to find the experts disagreeing on these highly salient issues. A tentative explanation could be that the experts disagreed about whether the Polish government acted sincerely or strategically with respect to the salient issues.

More generally, we find a significant and positive correlation between the proportion of salient issues and inter-expert coherence. The correlation between proportion of salient issues and percent agreement is .42 (significance level: .04). The correlation with the α for all issues equals .44 (significance level: .03). Limiting ourselves to the salient issues, the correlation with percent agreement becomes .43 (significance level: .03), and the correlation with the α is .56 (significance level: .004). If, for a given political actor, more

issues are salient, the experts are more in agreement. This holds not only for the salient issues but for all issues. If the European Constitution is considered more important in general (and not just for a few specific issues), there is more agreement among the experts about the national policy positions.

Conclusion

The DOSEI project shows that expert interview is an attractive data collection method, which has allowed the project to collect a wealth of (highly timely) information on all the relevant policy positions of all the main actors with respect to the draft European Constitution. However, rather obviously, the validity of the information collected by means of expert interviews crucially depends on the quality of the experts. By considering the reliability of the experts in the DOSEI project, we have applied these ideas in a political framework. A second, but equally important, objective has been to evaluate the quality of the DOSEI data.

Any theoretical link between the reliability and validity of data cannot simply be assumed. However, the Condorcet Jury Theorem can be used to argue for the existence of such a link. More coherent, i.e. reliable, experts are also more likely to be right, i.e. provide valid information, under some quite general and reasonable assumptions. The Condorcet Jury Theorem does not require all experts to be equally knowledgeable and they may be better informed on some issues rather than others.

Even though – or perhaps precisely because – a wide variety of indices exist for inter-coder reliability, there is no agreement about what is the 'best', most generally applicable statistic. Moreover, none of the existing indices is especially designed to evaluate the reliability of multiple experts. To provide an example, it is typical in content analysis that a small number of coders evaluate a large number of units (texts/sentences) on a small number of characteristics. In contrast, in political science, typically, a large number of experts (but only a few for each country) evaluate a small number of units (political actors) on a large number of characteristics (policy positions). It has been beyond the scope of this article to devise reliability tests that are most appropriate for expert interviews. Instead, we have presented the results for the most commonly used indices of inter-coder reliability.

In most cases, the levels of inter-expert agreement are acceptable, given the above-mentioned limitations. Generally, expert agreement approaches or meets the criteria set for the κ and α statistics. There are no large differences in expert agreement depending on the inclusion or exclusion of the 'leading'

experts. Experts are clearly more in agreement on salient issues. If more issues are considered to be salient for a particular country, the experts are more in agreement about the country's national position on all issues.

High-quality data are essential for empirical research. Not even the most sophisticated estimation procedure can yield valid conclusions on the bases of seriously flawed data. In principle, expert interviews can be used to collect data of high quality, but only if the experts are willing to go along. The reliability of expert data should thus be checked as a matter of routine. Even though reliability does not *guarantee* validity, it makes it more likely that a valid conclusion will be reached. At the same time, reliability should not be pursued at all costs: there is always a possibility that one expert is 'right' and all others are simply 'wrong'.

Notes

Earlier versions of this article were presented at the DOSEI Measurement conference, Speyer, 13 June 2004. We thank all participants for their helpful comments. Further, we thank the research assistants in the DOSEI project for their work on collecting and compiling the data. Finally, we are especially grateful to Professor Klaus Krippendorff for advising us on the use and interpretation of his α statistic.

1 Methodological and practical issues of expert interviews as data collection method are discussed in Richards (1996), Devine (2002: 197–215) and Burnham et al. (2004: 205–20).

2 For examples of comparative studies using expert interviews, see Bueno de Mesquita and Stokman (1994), Hooghe (1999), Bailer (2004) and Arregui et al. (2004). The use of expert interviews in EU policy-specific studies is commonplace; for examples of landmark studies, see Moravcsik (1998) and Dyson and Featherstone (1999).

3 The Jury Theorem can be traced back to the work of Condorcet (1785) and was 'rediscovered' by Black (1958); see also Grofman et al. (1983). Nurmi (2002: 49–59) provides a concise summary and discussion of recent research on the Condorcet Jury Theorem.

4 Di Eugenio and Glass (2004: 5) demonstrate that this applies to the Cohen's κ but not to the Fleiss–Siegel and Castellan extensions of the κ. We therefore report only the latter κ's. We further report and compare the κ with percent agreement and Krippendorff α.

5 The distributions of κ and α can be calculated only under limited conditions that, unfortunately, do not apply to our data. Consequently, we were unable to test whether the differences between the indices are statistically significant.

6 Given the complexity of the negotiations on the European constitutional Treaty and thus the questionnaire, a possible shortcoming of this approach was that there may not have been one single 'leading' expert. Instead, the quality of the information provided by the various experts may have depended on the specific issue area.

7 The results for the κ statistic are nearly identical and are available on request from the authors.

References

Arregui, Javier, Frans N. Stokman and Robert Thomson (2004) 'Bargaining in the European Union and Shifts in Actors' Policy Position', *European Union Politics* 5(1): 47–72.

Austin-Smith, David and Jeffrey S. Banks (1996) 'Information Aggregation, Rationality, and the Condorcet Jury Theorem', *American Political Science Review* 90(1): 34–45.

Bailer, Stefanie (2004) 'Bargaining Success in the European Union: The Impact of Exogenous and Endogenous Power Resources', *European Union Politics* 5(1): 99–123.

Ben-Yashar, Ruth and Jacob Paroush (2000) 'A Nonasymptotic Condorcet Jury Theorem', *Social Choice and Welfare* 17: 189–99.

Black, Duncan (1958) *The Theory of Committees and Elections*. Cambridge: Cambridge University Press.

Bueno de Mesquita, Bruce and Frans N. Stokman (eds) (1994) *European Community Decision Making: Models, Applications, and Comparisons*. New Haven, CT: Yale University Press.

Burnham, Peter, Karin Gilland, Wyn Grant and Zig Layton-Henry (2004) *Research Methods in Politics*. Houndmills: Palgrave.

Carletta, Jean (1996) 'Assessing Agreement on Classification Tasks: The Kappa Statistic', *Computational Linguistics* 22(2): 249–54.

Cohen, Jacob (1960) 'A Coefficient of Agreement for Nominal Scales', *Educational and Psychological Measurement* 20: 37–46.

Condorcet, Marquis de, M.J.A.N. de Caritat (1785) 'An Essay on the Application of Analysis to the Probability of Decisions Rendered by a Plurality of Votes'; in *Classics of Social Choice*, ed. and transl. Iain McLean and Arnold B. Urken, pp. 91–112. Ann Arbor: University of Michigan Press, 1995.

Devine, Fiona (2002) 'Qualitative Methods', in David Marsh and Gerry Stoker (eds) *Theory and Methods in Political Science*, 2nd edn, pp. 197–215. Houndmills: Palgrave.

Di Eugenio, Barbara and Michael Glass (2004) 'The Kappa Statistic: A Second Look', *Computational Linguistics* 30(1): 95–101.

Dyson, Kenneth and Kevin Featherstone (1999) *The Road to Maastricht: Negotiating Economic and Monetary Union*. Oxford: Oxford University Press.

Flash Eurobarometer (2004) 'The Future European Constitution', Eurobarometer 159, TNS Sofres/EOS Gallup Europe, January.

Fleiss, Joseph L. (1971) 'Measuring Nominal Scale Agreement among Many Raters', *Psychological Bulletin* 76: 378–82.

Grofman, Bernard, Guillermo Owen and Scott L. Feld (1983) 'Thirteen Theorems in Search of the Truth', *Theory and Decision* 15: 261–78.

Holsti, Ole R. (1969) *Content Analysis for the Social Sciences and Humanities*. Reading, MA: Addison-Wesley.

Hooghe, Liesbet (1999) 'Supranational Activists or Intergovernmental Agents? Explaining the Orientations of Senior Commission Officials toward European Integration', *Comparative Political Studies* 32(4): 435–63.
Hughes, Marie A. and Dennis E. Garrett (1990) 'Intercoder Reliability Estimation-Approaches in Marketing: A Generalizability Theory Framework for Quantitative Data', *Journal of Marketing Research* 27: 185–95.
King, Gary, Robert O. Keohane and Sidney Verba (1994) *Designing Social Inquiry. Scientific Inference in Qualitative Research.* Princeton, NJ: Princeton University Press.
Kolbe, Richard H. and Melissa S. Burnett (1991) 'Content-Analysis Research: An Examination of Application with Directives for Improving Research Reliability and Objectivity', *Journal of Consumer Research* 18: 243–50.
Krippendorff, Klaus (2004a) 'Reliability in Content Analysis: Some Common Misconceptions and Recommendations', *Human Communication Research* 30(3): 411–33.
Krippendorff, Klaus (2004b) *Content Analysis: An Introduction to Its Methodology*, 2nd edn. Thousand Oaks, CA: Sage.
Lombard, Matthew, Jennifer Snyder-Duch and Cheryl Campanella Bracken (2004) 'Content Analysis in Mass Communication: Assessment and Reporting of Intercoder Reliability', *Human Communication Research* 28(4): 587–604.
Moravcsik, Andrew (1998) *The Choice for Europe: Social Purpose and State Power from Messina to Maastricht*. London: UCL Press.
Neuendorf, Kimberly A. (2002) *The Content Analysis Guidebook.* Thousand Oaks, CA: Sage.
Nitzan, Shmuel and Jacob Paroush (1982) 'Optimal Decision Rules in Uncertain Dichotomous Choice Situation', *International Economic Review* 23: 289–97.
Nurmi, Hannu (2002) *Voting Procedures under Uncertainty*. Berlin: Springer.
Owen, Guillermo, Bernard Grofman and Scott L. Feld (1989) 'Proving a Distribution-Free Generalization of the Condorcet Jury Theorem', *Mathematical Social Sciences* 17: 1–16.
Richards, David (1996) 'Elite Interviewing: Approaches and Pitfalls', *Politics* 16(3): 199–204.
Riffe, Daniel and Alan A. Freitag (1997) 'A Content Analysis of Content Analyses: Twenty-five Years of *Journalism & Mass Communication Quarterly*', *Journalism & Mass Communication Quarterly* 74(4): 873–82.
Scott, William A. (1955) 'Reliability of Content Analysis: The Case of Nominal Scale Coding', *Public Opinion Quarterly* 17: 321–5.
Siegel, Sidney and N. John Castellan Jr (1988) *Nonparametric Statistics for the Behavioral Sciences*, 2nd edn. New York: McGraw-Hill.

About the authors

Han Dorussen is Senior Lecturer in the Department of Government, University of Essex, Wivenhoe Park, Colchester CO3 4SQ, Essex, UK.
Fax: +44 1206 873 234
E-mail: hdorus@essex.ac.uk

Hartmut Lenz is a PhD candidate in the Department of Government, University of Essex, Wivenhoe Park, Colchester CO3 4SQ, Essex, UK.
Fax: +44 1206 873 234
E-mail: hlenz@essex.ac.uk

Spyros Blavoukos is a PhD candidate in the Department of Government, University of Essex, Wivenhoe Park, Colchester CO3 4SQ, Essex, UK.
Fax: +44 1206 873 234
E-mail: sblavo@aueb.gr